Measuring morphological complexity & information density

Yoon Mi OH (yoonmioh@ajou.ac.kr)

1. Morphological complexity (labelled as **O_MC** in the template)

The measure of morphological complexity is adopted from the methodology proposed in Lupyan and Dale (2010). In their paper, 28 linguistic features are chosen to account for the inflectional morphology from WALS. The score of morphological complexity is calculated by dichotomically distinguishing between lexical and inflectional coding strategies and summing assigned values (-1 for lexical and 0 for morphological strategies) to the 29 linguistic features displayed in Table 1.

Table 1: Measure of morphological complexity. Features are chosen and classified following (Lupyan and Dale, 2010) with descriptions taken from WALS (Dryer & Haspelmath, 2013). Two features, Definite articles (37A) and Indefinite articles (38A), are considered together as one linguistic feature in (Lupyan and Dale, 2010) but are separately taken into account in the present method.

Feature (WALS code)	**Description**
Morphological type	
Fusion of selected inflectional formatives (20A)	The degree to which grammatical markers (formatives) are phonologically connected to a host word or stem
Prefixing vs. suffixing (26A)	The degree to which languages use prefixes or suffixes in their inflectional morphology
Cases	
Number of cases (49A)	The number of case categories represented in a language's inflectional system
Case syncretism (28A)	The ways in which a single inflected form represents two or more case functions
Alignment of case marking of full noun phrases (98A)	The ways in which core argument noun phrases are marked to indicate which particular core argument position they occupy
Verb morphology	
Inflectional synthesis of the verb (22A)	The strategies of expressing grammatical categories either by individual words or by affixes attached to some other words
Alignment of verbal person marking (100A)	The ways in which the two arguments of the transitive verb align with the sole argument of the intransitive verb
Agreement	
Person marking on verbs (102A)	The number and identity of the arguments of a transitive clause which display person marking on the verb
Person marking on adpositions (48A)	The strategies of person marking used to relate an object to another nominal or verbal constituent on the

	basis of a more or less specific semantic relationship
Syncretism in verbal person/number marking (29A)	The ways in which multiple person values underlie a single form in the inflectional marking of subject person in verbs
Possibility and evidentials	
Situational possibility (74A)	The strategies used to express situational possibility in positive main clauses
Epistemic possibility (75A)	The strategies used to express epistemic possibility in positive main clauses
Overlap between situational and epistemic modal marking (76A)	The extent to which languages have identical markers for situational and epistemic modality
Semantic distinctions of evidentiality (77A)	The presence of grammatical markers of evidentiality which express the evidence a speaker has for his/her statement
Negation, plurality, interrogatives	
Negative morphemes (112A)	The nature of morphemes signaling clausal negation in declarative sentences
Occurrence of nominal plurality (34A)	The extent to which plural markers on full nouns are used in a language
Associative plural (36A)	It consists of a noun X and some other materials referring to 'X and other people associated with X'
Position of polar question particles (92A)	The position of question particles in polar questions (questions that elicit the equivalent of a 'yes' or 'no' response)
Tense, possession, aspect, mood	
Future tense (67A)	The distinction between languages with and without inflectional marking of future time reference
Past tense (66A)	The ways in which past/non-past distinction is marked grammatically
Perfective/Imperfective aspect (65A)	The distinction between languages with and without the perfective/imperfective grammatical marking
Morphological imperative (70A)	The extent to which languages have second person singular and plural imperatives as dedicated morphological categories
Position of pronominal possessive affixes (57A)	The distinction between languages with and without possessive suffixes and prefixes on noun
Possessive classification (59A)	The forms of possessive marking whose choice is conditioned lexically by the possessed noun
Optative (73A)	An inflected verb form dedicated to the expression of the wish of the speaker
Articles, demonstratives, pronouns	
Definite articles (37A)	A morpheme which accompanies nouns and codes definiteness or specificity

Indefinite articles (38A)	A morpheme which accompanies a noun and signals that the noun phrase denotes something not known to the hearer
Distance contrasts in demonstratives (41A)	The ways in which deictic expressions indicating the relative distance of a referent in the speech situation vis-à-vis the deictic center are marked
Expression of pronominal subjects (101A)	The ways in which a pronominal subject is expressed by a morpheme or morphemes coding semantic or grammatical features of the subject

The relevant information for each linguistic feature is solely obtained from the information provided in WALS. However, WALS does not provide all the information regarding the features presented in the table above. Therefore, the score was calculated by dividing the overall score by the total number of available linguistic features.

The features are distinguished into two types of variables: metric (quantitative) and non-metric (categorical or qualitative) variables. The present method differs from the measure used in (Lupyan & Dale, 2010) in a way that the latter converts non-metric, categorical variables with multiple values into dichotomous variables by assigning two possible values for each feature, -1 for lexical and 0 for inflectional morphological strategy. On the contrary, in our method, some features are considered as continuous variables. For instance, to reflect the quantitative variables, such as the number of case categories (49A) and the number of grammatical categories expressed by the inflectional synthesis of the verb (22A), all the values are normalized between 0 and -1, including those attributed to continuous variables. Taking normalized values of continuous variables into account is assumed to better represent the degree of morphological complexity since they specify the evaluation criteria.

The methods for quantifying linguistic complexity differ as a function of linguistic module in question: bottom-up or usage-based approach. In (Dahl, 2004), the author distinguishes two notions of linguistic complexity. The first notion of linguistic complexity regards language as a system (*system complexity*) and measures the richness of a system in terms of its resources. The second notion applies to the structure of expressions (*structural complexity*). Such distinction of linguistic complexity accounts for the differences between the methodologies used to measure morphological and phonological complexities. According to Dahl, system complexity could be measured at the phonological level and structural complexity could be calculated at the morphological level of analysis, but not exclusively. Since morphology investigates the structure and form of words, it should be crucial to take morphological coding strategies, i.e. structural complexity, into account for measuring morphological complexity. The present method thus calculates morphological complexity using both metric and non-metric variables from the traditional grammar-based method.

2. Information density (labelled as **O_SID** (syllable information density) and **O_WID** (word information density) and in the template)

A syllable information density, SID, refers to the average amount of information conveyed per syllable (*SI*) and is calculated for each target language as the average ratio between the total number of syllables in a sentence in Korean and the number of syllables of the corresponding sentence translated in the target language (Pellegrino et al., 2011; Oh, 2015). The average amount of information per syllable (SI_L^s) is defined as the division of the semantic content *C* of sentence *s* in language *L* (C_L^s) by the number of its syllables (σ_L^s).

$$SI_L^s = \frac{C_L^s}{\sigma_L^s}$$

To compute SID, automatic syllabification tool is required to syllabify the data in each language. For this workshop, only the data in 10 languages (Basque, German, English, Finnish, French, Georgian, Korean, Russian, Spanish, and Turkish) are syllabified using currently available automatic syllabification tools (Reichel & Kisler, 2014; Oh, 2015). Among 10 languages, Korean is chosen as a reference since it has the lowest score of morphological complexity (in other words, it uses more lexical strategies and is hypothetically more phonologically complex than the other nine languages). As the fully Parallelized Bible Corpus (Track A) is used, the semantic content of each sentence is assumed to be equivalent for all languages ($C_L^s = C_{KOR}^s$).[1] SID is computed by a pairwise comparison of the number of syllables of sentence *s* in Korean (σ_{KOR}^s) and in a target language (σ_L^s).

$$SID_L = \frac{1}{S}\sum_{s=1}^{S}\frac{SI_L^S}{SI_{KOR}^s} = \frac{1}{S}\sum_{s=1}^{S}\frac{C_L^s}{\sigma_L^s} \times \frac{\sigma_{KOR}^s}{C_{KOR}^s} = \frac{1}{S}\sum_{s=1}^{S}\frac{\sigma_{KOR}^s}{\sigma_L^s}$$

A word information density, WID, corresponds to the average amount of information conveyed per word (*WI*) and is defined as the division of the semantic content *C* of sentence *s* in language *L* (C_L^s) by the number of its words (ω_L^s). Among 48 languages of the fully Parallelized Bible Corpus (Track A), Thai is the most isolating language with the lowest morphological complexity and therefore is used as an external reference. Since $C_L^s = C_{THA}^s$, WID is computed by a pairwise comparison of the number of words of sentence *s* in Thai (ω_{THA}^s) and in a target language (ω_L^s).

$$WI_L^s = \frac{C_L^s}{\omega_L^s}$$

[1] This measure can be only applied to the data which is carefully designed to minimize and avoid any variation with respect to the translation.

$$WID_L = \frac{1}{S}\sum_{s=1}^{S}\frac{WI_L^S}{WI_{THA}^s} = \frac{1}{S}\sum_{s=1}^{S}\frac{C_L^s}{\omega_L^s} \times \frac{\omega_{THA}^s}{C_{THA}^s} = \frac{1}{S}\sum_{s=1}^{S}\frac{\omega_{THA}^s}{\omega_L^s}$$

This syntagmatic measure of information density computes the average amount of information carried by syllables and words at the sentence level on the local scale and differs from studies related to the principle of Uniform Information Density (Frank and Jaeger, 2008) which deals with the variation of information transmitted during communication on the global scale, which requires large data.

Dryer, M. S. & Haspelmath, M. (Eds.) (2013). *The World Atlas of Language Structures Online.* Leipzig: Max Planck Institute for Evolutionary Anthropology.

Frank, A. & Jaeger, T. F. (2008). Speaking rationally: Uniform information density as an optimal strategy for language production. In *Proc. of the 30th annual meeting of the cognitive science society*, Washington, DC: Cognitive Science Society. pp. 933-938.

Lupyan, G. & Dale, R. (2010). Language Structure Is Partly Determined by Social Structure. *PLoS ONE 5*(1): e8559.

Oh, Y. (2015). *Linguistic complexity and information: quantitative approaches*, PhD thesis, Language Science, University of Lyon 2.

Pellegrino, F., Coupé, C., & Marsico, E. (2011). Across-language perspective on speech information rate. *Language 87*(3), 539–558.

Reichel, U.D., & Kisler, T. (2014). Language-independent grapheme-phoneme conversion and word stress assignment as a web service. In: Hoffmann, R. (Ed.): *Elektronische Sprachverarbeitung. Studientexte zur Sprachkommunikation 71*, pp42-49, TUDpress, Dresden.